Automatic Generation of Questions from DBpedia

Oscar Rodríguez Rocha, Catherine Faron Zucker, Alain Giboin, Aurélie Lagarrigue

Automatic Generation of Questions from DBpedia

Oscar Rodríguez Rocha*, Catherine Faron Zucker, Alain Giboin, Aurélie Lagarrigue

Université Côte d'Azur, CNRS, Inria, I3S, France
E-mail: oscar.rodriguez-rocha@inria.fr
E-mail: faron@unice.fr
E-mail: alain.giboin@inria.fr
E-mail: aurelie.lagarrigue@inria.fr
*Corresponding author

Abstract:
The production of Educational quizzes is a time-consuming task that can be automated by taking advantage of existing knowledge bases available on the Web of Linked Open Data (LOD) such as DBpedia. For these quizzes to be useful to learners, their questions must be generated in compliance with the knowledge and skills necessary to master for each subject and school year according to the official educational standards. We present an approach that exploits structured knowledge bases that contain the knowledge and skills, from which it selects a set of DBpedia resources relevant to a specific school subject and year. This set of resources is enriched with additional related resources, through the NEFCE heuristic. Finally, question generation strategies are applied to the graph generated with this set of resulting resources. Likewise, we provide an evaluation of two knowledge bases and of the proposed NEFCE heuristics.

Keywords: eEducation; Semantic Web; Linked Data.

Biographical notes: Oscar Rodríguez Rocha is a researcher at the French Institute for Research in Computer Science and Automation (Inria). He received his Ph.D. in Computer and Control Engineering from the Polytechnic University of Turin in 2014. He works for Wimmics, a joint research team between Inria, CNRS and Université Nice Sophia Antipolis. His main research interests are Knowledge Representation and Reasoning, Recommender Systems, and Semantic Web. Currently he focuses on the automatic generation of educational quizzes from knowledge graphs.

Catherine Faron Zucker is an Associate Professor of Computer Science at Université Nice Sophia Antipolis since 2002. She received her PhD in Computer Science from Université Paris 6. She is vice-head of the Wimmics joint research team between Inria, CNRS and Université Nice Sophia Antipolis. Her research activity focuses on Knowledge Engineering and Modeling, Graph based Knowledge Representation and Reasoning, Ontologies, Semantic Web and Social Web. Her main application domains are Collective Memories and Intelligent Tutoring Systems.

Alain Giboin is a senior researcher at the French Institute for Research in Computer Science and Automation (Inria). He is a member of the Wimmics joint research team between Inria, CNRS and Université Nice Sophia. His research activity focuses on Human Computer Interactions (HCI).

Aurélie Lagarrigue Aurelie Lagarrigue is an Instructional Designer at the French Institute for Research in Computer Science and Automation (Inria). She received

her PhD in Cognitive Neurosciences from Mirail University of Toulouse. She is interested in learning and docimology. She currently works at the Inria's Learning Lab in the creation of learning modules and assessments for online courses.

1 Introduction

Educational quizzes are a popular and effective instrument that enables students or learners to informally and entertainingly discover and/or deepen their knowledge about a certain topic. Although it has been shown that quizzes can be automatically generated using Semantic Web technologies and exploiting Linked Open Data (LOD), so far there is no approach that achieves that the questions of these quizzes are relevant according to a domain or subject and that also correspond to the educational level of each student.

In the research work presented in this paper, we have considered the use of referentials of knowledge and skills to select DBpedia resources relevant to specific topics and school courses in France and from which, the questions can be automatically generated. For this, we have carried out an evaluation of two existing referentials of knowledge and skills to determine which one allows a better selection of resources with these characteristics.

Finally, we describe our approach to generate questions based on DBpedia resources selected from the chosen referential and enriched with additional related resources through the use of a proposed heuristics (*NEFCE*), which results in a greater number of relevant questions being generated.

Concretely, the present work contributes to answer the following research questions:

1. How to select resources from DBpedia that are relevant to a specific subject and to a specific school level?
2. Which referential of knowledge and skills is the most appropriate to use for the selection of suitable resources from DBpedia?
3. Which approach to use for the generation of questions about a specific subject, that takes into account the school level of a student?

This paper is structured as follows: In section 2, we present the related works. In the section 3, we describe two structured knowledge bases that can be used as a reference of the knowledge and skills required in each course of the French educational system. In the section 4 we detail the process to extract a knowledge graph from DBpedia according to the referentials of knowledge and skills and how this graph can be enriched using the *NEFCE* heuristics. In section 5, we detail our approach to automatically generate of questions from knowledge graphs. In the section 6, we describe how we evaluated the resources contained in the generated knowledge graphs and also the questions generated from them. Section 7 contains our conclusions and future works.

2 Related Work

We have organized the existing related works presented in this section, according to the different steps to which they correspond in our approach. In this way, we first present some

works on educational ontologies (related to Section 3). From the analysis of these works, we have been able to select an ontological model to be used as an educational reference for the generation of questions in our approach. Then we present existing works on knowledge extraction from DBpedia (related to the step of our approach described in Section 4). Finally we present existing works on automatic question generation (related to the step of our approach described in Section 5).

2.1 Related works on educational ontologies

Many ontologies have been developed in order to describe and represent different aspects of education. A very complete literature survey about the development and use of ontologies in e-learning systems (Yahya et al. [2015]), classifies the literature found in four categories: (1) curriculum management and modeling, (2) description of learning domains, description of learners' data and (4) description of e-Learning services. Within the first category, CURONTO (Al-Yahya et al. [2013]) is an ontological model designed for the management of a curriculum, allowing also to simplify its review and assessment.

Gescur (Dexter and Davies [2009]) is a tool dedicated to the management and evaluation of the implementation of a curriculum, which mainly facilitates the curriculum management process. This tool mainly relies on an ontology of concepts relevant to curriculum management in Secondary Schools, such as teachers, departments, objectives, subjects, modules, tasks, documents, policies, activities, learning objects, quality criteria, etc.

An educational semantic web ontology is proposed in (Bucos et al. [2010]), focusing on representing higher education concepts and assisting specialized e-learning systems .

OntoEdu (Guangzuo et al. [2004]) is an educational platform architecture for e-Learning that relies on an activity ontology that describes all the educational activities and the relations among them, and a material ontology which describes the educational content organization, to allow to discover automatically, invoke, monitor and compose learning paths.

To the best of our knowledge, none of the above ontologies focuses on the representation and description of knowledge and skills that the students should acquire progressively and none of them has been populated with an existing standard. For this reason, the approach described in this paper is based on EduProgression (Rodríguez Rocha et al. [2017]), an ontological model for educational progressions, that has been formalized in OWL, and is based on the modeling of the knowledge and skills of an educational system according to its different educational levels or school years. Likewise this ontology is currently populated with the official progressions of the French educational system for the subjects of History, Geography and Experimental Sciences and Technology.

2.2 Related work on Knowledge extraction from DBpedia

A related approach (Lalithsena et al. [2016]) focuses on identifying "a minimal domain-specific subgraph by utilizing statistics and semantic-based metrics". It targets DBpedia as a knowledge base and focuses on identifying entities and relationships strongly associated with a domain of interest. The domain of interest is initially identified through a representative entity. For example, for a movie, the corresponding entity could be `dbo:Film`, and the resources having this type (`rdf:type`) in DBpedia, will be the entities that describe the domain. This is the biggest difference with respect to our approach, we consider that in some cases DBpedia classes can be very general, and therefore, as our algorithm is not based on the identification of resources based on their class, it allows to define domains in a more granular and specific way.

Other related works (Mirizzi et al. [2010a], Mirizzi et al. [2010b] and Mirizzi et al. [2010c]) propose to exploit semantic relations stored in DBpedia to extract and rank resources related to the user context given by the keywords she enters in a search engine to formulate her query, with the aim of improving the selection of ads relevant to the user context, to be added in the results of the (sponsored) search engine. When compared to this work, in order to extract DBpedia resources semantically related to an input text or set of terms describing the targeted domain for the generation of quizzes, our approach relies only on DBpedia and does not consider external sources of unstructured knowledge. Furthermore, the estimation of the *strength of the connection* between two resources linked through a *wikilink* property is not yet considered.

Finally, we can also find related works that deal with resource discovery and graph exploration. A very complete systematic literature review on Recommendation Systems based on Linked Data (Figueroa et al. [2015]) lists the state of the art works in this area. None of the mentioned works focuses directly on recovering DBpedia resources corresponding to a subject or year of an educational reference. However, these works were a source of inspiration to formulate our *NEFCE* heuristics.

2.3 Related works on automatic generation of questions

An approach for automatically generating computer-assisted assessments (CAA) from Semantic Web-based domain ontologies is proposed in (Cubric and Tosic [2010]). It consists in the addition of annotations to the meta-ontology used for the generation of questions and the addition of a semantic interpretation of the mapping between the *domain ontology* and the target *question ontology*. The semantic interpretation is based on the notion of *question templates* in the Bloom's taxonomy of educational objectives. However, differently from our approach, this work does not take into account the generation of questions according to the required knowledge and skills required in each educational level.

The OntoQue system (Al-Yahya [2011]) is dedicated to the generation of objective assessment items based on domain ontologies. It uses knowledge inherent in the ontology about classes, properties, and individuals to generate semantically correct assessment items. Its authors evaluated the system with four OWL ontologies from different domains. The major limit of this approach, is that the questions are generated only from simple triples of the knowledge base, and are based on manually written predicates.

In order to generate educational assessment items using Linked Open Data (DBpedia), an approach is proposed in (Foulonneau [2012]) consisting in a streamline to create variables and populate simple choice item models using the IMS-QTI standard. A set of specific categories common to a set of resources is statically selected, from which questions, answers and distractors will be generated using SPARQL queries. When compared to this work, our approach is based on an educational reference and aims at the automatic generation of questions, answers and distractors.

An approach for the validation of ontologies which abstracts the complexity of formal languages by using a set of questions and answers to which the expert is subjected has been proposed in (Abacha et al. [2016]). Its main purpose is to validate the conceptualization of the domain. Automatic reasoning and verbalization techniques are used to transform the facts present in an ontology into questions expressed in natural language, to be evaluated by a domain expert. The answers to the generated questions are then processed to automatically validate or correct the ontology. When compared to this work, our approach does not aim and has not been evaluated for the validation of domain ontologies.

Automatic generation of Multiple Choice Questions (MCQs), is also possible through an unsupervised Relation Extraction technique used in Natural Language Processing (Afzal and Mitkov [2014]). This approach aims to identify the most important named entities and terminology in a document and then recognize semantic relations between them, without any prior knowledge as to the semantic types of the relations or their specific linguistic realization. In contrast, our approach relies on educational references and in the knowledge present in DBpedia for the automatic generation of questions.

Another related work aims the automatic generation of Multiple Choice Questions (MCQs) from OWL ontologies (Alsubait et al. [2014, 2015, 2016]). This approach is based on the structure of a knowledge base to generate multiple choice questions by exploiting the existing relationships between the classes. The main limitation of this approach is that it only considers simple or direct relationships between classes for the generation of questions.

A set of strategies for the automatic generation of questions, answers and distractors from a given ontology has also been proposed in (Papasalouros et al. [2008]). Such strategies are classified into three categories: class-based, property-based and terminology-based. An extension of this work (Rodriguez Rocha and Faron Zucker [2017]) proposes a new classification of such strategies to generate quizzes and apply them to DBpedia. As it will be explained in the following, our approach of automatic generation of quizzes presented here is a continuation of this work.

3 Educational Referentials of Knowledge and Skills

The main purpose of our approach to generating questions from DBpedia, is to propose to the students the questions that correspond to the criteria defined by education and domain experts. This is the reason why we have considered and evaluated the use of the following educational references.

3.1 EduProgression

EduProgression (Rodríguez Rocha et al. [2017]) is an ontological model formalized in the standard Ontology Web Language (OWL) to represent educational progressions or programs as defined in the French common bases of 2006 [1] and 2016. In this model, we identified the following main classes:

- **Element of Knowledge and Skills (EKS).** As knowledge and skills are the keystones of the common bases, this element is the key concept of our model. It is represented by the `EKS` class, which is the main class of the *EduProgression* ontology. This class is common for the two releases of the common base.

  An element of knowledge and skills is associated to a set of knowledge pieces (class `Knowledge`) and/or skills (class `Skill`) for a specific French school cycle (class `Cycle`) or course (class `Course`) that may contain reference points (class `PointOfReference`) and also a vocabulary of associated terms (class `Vocabulary`). More precisely:

  - `Knowledge`. Instances of class `Knowledge` are also instances of `skos:Concept` and each one belongs to a `skos:ConceptScheme` that contains all the knowledge pieces of a given progression. An instance of `EKS` is related to an instance of `Knowledge` through property `hasKnowledge`.

- `Skill`. An instance of `EKS` is related to an instance of `Skill` through property `hasSkill`.
- `Course`. In the French common base of 2006, the skills that students are expected to develop are defined by cycle, and each cycle is organized into courses. For example, "the consolidation cycle" includes "the second year elementary course" (CE2), "the first-year intermediate course" (CM1) and "the second year intermediate course" (CM2). In this context, an instance of class `Course` represents a course in a cycle. An instance of `EKS` is related to an instance of `Course` through property `hasCourse`.
- `Cycle`. In the French common base of 2016, the skills that students are expected to develop are defined only by cycle. An instance of `EKS` is related to an instance of `Cycle` through property `hasCycle`.
- `PointOfReference`. An instance of class `PointOfReference` represents an educational reference element on a specific element of knowledge and skills (an instance of `EKS`). An instance of `EKS` is related to an instance of class `PointOfReference` through property `hasPointOfReference`.
- `VocabularyItem`. Each element of knowledge and skills has vocabulary items. This vocabulary is compatible for the two common bases. An instance of `EKS` is related to an instance of class `VocabularyItem` through property `hasVocabularyItem`. A vocabulary item is also an instance of `skos:Concept` and it is related to an instance of `skos:ConceptScheme` which gathers the concepts of the thesaurus of the progression.

- **Progression.** In the common base released in 2006, the progressive acquisition of knowledge and skills is defined as a "progression", while for the current common base, the progressive acquisition of skills is defined as a "program". Therefore, a progression or a program in our model, is represented as an instance of the class `Progression`. It can be associated to an existing learning domain (through property `hasLearningDomain`) and to one or several EKSs (through property `hasEKS`).
- **Learning domain.** A learning domain represents a school subject like *History* or *Mathematics*. The learning domain is represented, in the ontology EduProgression, by an instance of the class `LearningDomain`, and it is also an instance of `skos:Concept` that is part of (only) one `skos:ConceptScheme` containing the only learning domains of a progression. Also, as they are SKOS concepts, learning domains are hierarchically organized by using the `skos:broader` and/or `skos:narrower` properties. A learning domain can be associated to a `Progression` or an `EKS`.
- **Skills Domain.** For the common base of 2016, each domain of skills can be represented in our model by an instance of the class `SkillsDomain`. An instance of `EKS` can be associated to one or many instances of `SkillsDomain` through property `hasSkillsDomain` to represent the skills of the domain(s) that it targets.

The EduProgression ontology, which is freely accessible online[2], is composed of 7 OWL classes and 8 OWL object properties and it has been populated with the progressions of the French educational system for the subjects of History, Geography and Experimental Sciences and Technology and for the French school courses *CE2*, *CM1* and *CM2*. Below is an excerpt of an EKS described through EduProgression.

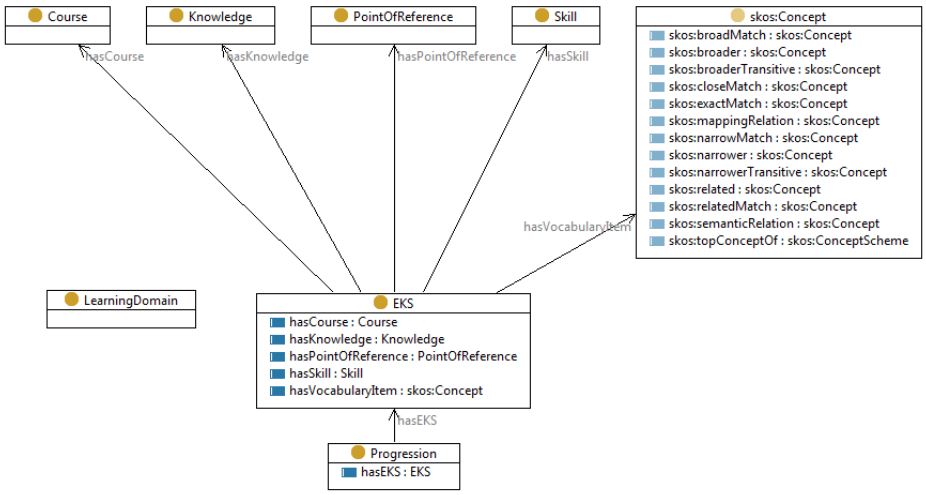


Figure 1: Classes and Properties of the EduProgression Ontology

```
@prefix : <http://ns.inria.fr/semed/eduprogression#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
:Lire_une_carte a :EKS ;
    rdfs:label "Read a map"@en ;
    rdfs:label "Lire une carte"@fr ;
    :hasCourse :CM2 ;
    :hasLearningDomain :Capacites_propres_a_la_geographie ;
    :hasSkill :Realiser_une_carte_simple,
        :Utiliser_la_legende_dune_carte .
:CM2 a :Course ;
    rdfs:label "Middle course second year"@en ;
    rdfs:label "Cours moyen 2e annÃ©e"@fr  .
:Capacites_propres_a_la_geographie a :LearningDomain,
        skos:Concept ;
    skos:prefLabel "Capacities specific to the geography"@en ;
    skos:broader :Elements_de_connaissances_et_de_competences_Geographie ;
    skos:inScheme :GeographieLearningDomains .
:Realiser_une_carte_simple a :Skill ;
    rdfs:label "Create a simple map"@en .
:Utiliser_la_legende_dune_carte a :Skill ;
    rdfs:label "Use the legend of a map"@en .
```

3.2 Les Incollables Knowledge Base

Intelliquiz is a collaborative project between Inria[3] and Qwant[4], that aimed to develop an automatic generation engine for quizzes made up of questions, their correct answer(s) and a set of distractors. Initially, quizzes could be generated taking as a reference the topics and the information currently present in the limited dataset of questions of *Les Incollables*[5], using the Web of Data as a source of knowledge.

Les Incollables is a world-famous questions game in French, targeted mainly to contribute to educate children and young people by playing. A large amount of the questions of the game are multiple-choice questions, which were created manually by domain experts and written on paper cards and several non-digital formats. These questions were generated with the different levels of the French education system in mind.

As part of this project, around 160,000 questions of the game belonging to 6 French school courses (*CP*, *CE1*, *CE2*, *CM1*, *CM2* and *6e*) have been subjected to a digitization process performed by the authors of this article and a group of 3 pedagogical engineers from Qwant. This semi-automatic process consisted of manually scanning game cards, documents and images containing the questions of the game. Then an automatic batch OCR (Optical Character Recognition)[6] process was applied to all those scanned files to output the questions in digital format. Then, another automatic text cleaning process was applied to the text of the questions to reduce possible errors. Finally, the resulting questions were anotated based on OWL ontologies and stored in a graph data format (RDF). Such OWL ontologies used were:

- *Linquest*[7]. An OWL vocabulary that describes and represents the most common elements of multiple-choice quizzes, such as a pair of a question and its answer `QA`, the natural language representation of such question `NLQuestion` and such answer `NLAnswer`, a set of questions `QASet` and set of sets of questions `QAMultiSet`.
- *Incollables*[8]. An OWL vocabulary that describes and represents the most specific aspects of the quizzes of the *Les Incollables* game, such as the theme (subject or subjects) of a Linquest set of questions (`QASet`) `hasTheme` and its level `hasLevel`. Also the DBpedia resource(s) related to a Linquest question (`QA`) `hasRelatedDBpediaResource`.
- *FrenchEdu*[9]. It formalizes the cycles, courses and related information of the French educational system, making possible to represent them as a formal and machine-understandable model. Its main classes are `Degree` and `School`.

An excerpt of the *Les Incollables Knowledge Base* is shown below.

```
@prefix dbpedia-fr: <http://fr.dbpedia.org/resource/> .
@prefix inc: <http://www.gaya-technology.com/incollables#> .
@prefix lq: <http://ns.inria.fr/semed/linquest#> .
@prefix fe: <http://ns.inria.fr/semed/frenchedu#> .
inc:QASet-HG-CE1 a lq:QASet ;
    lq:hasQA <#Q2163> ;
    inc:hasLevel fe:Cours_elementaire_premiere_annee ;
    inc:hasTheme inc:HistoireGeoEdCiv .
inc:Q2163 a lq:QA ;
    lq:NLQuestion "Quelle est la capitale de l'Espagne ?"@fr ;
    lq:NLQuestion "What is the capital of Spain ?"@en .
```

4 Extraction of a Knowledge Graph from DBpedia According to Referentials of Knowledge and Skills

The basis to automatically generate useful quizzes for the learners, is the selection of resources from a knowledge graph that are relevant to a specific subject and a school year

according to reference knowledge and skills. Considering that the knowledge bases on the Semantic Web use different ontologies and ways to structure their data, we decided to focus our study on DBpedia, which is one of the most used knowledge bases available on the Linked Open Data, and provides a large amount of resources from different domains. To extract a subgraph from DBpedia that contains knowledge from a specific subject and a school year, to which we refer to as a *knowledge graph*, our approach consists in the first three steps of Figure 2 (the remaining two steps will be detailed in section 5). These are as follows:

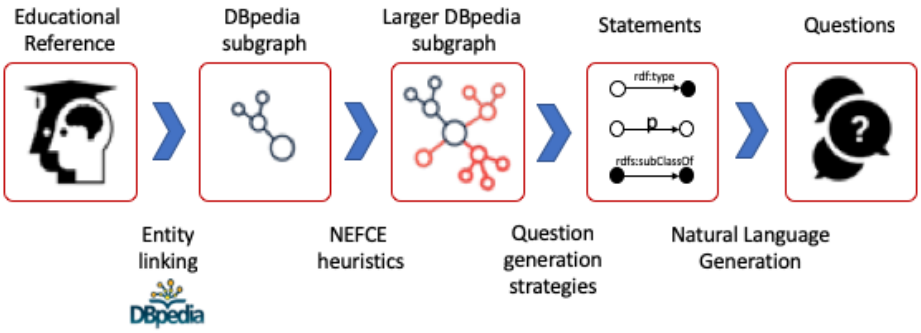


Figure 2: Overall process of automatic question generation from DBpedia according to Educational Referentials

- For a given subject and a school year, by means of a SPARQL query, we retrieve either the textual representations of the corresponding competences and skills (instances of `EKS`) in the case of *EduProgression*, or the textual representation of the corresponding questions (instances of `QA`) in the case of the knowledge base *Les Incollables*.
- An automatic entity-linking[10] process using DBpedia Spotlight[11], is applied to the obtained textual representations in order to output a set of related DBpedia resources *R*, that will be stored in a DBpedia subgraph (as shown in the first step of Figure 2).
- The *NEFCE* heuristics (detailed in Algorithm 1) is applied to the set of resources *R* stored in the DBpedia subgraph, in order to enrich it with the *wikilinks* (related DBpedia resources through property `dbpedia-owl:wikiPageWikiLink`) having relevant categories. For this, we build a set of categories *C* resulting from a dedicated SPARQL query on DBpedia searching for the value of property `dcterms:subject` or property path `dcterms:subject/skos:broader` for each resource in *R*. Then we restrict to the subset of *C* limited to the *k* categories having the most related resources (C_{topK}). The wikilinks of the resources in *R* which have a category in C_{topK} are added to *R*. The value of the number of categories *k* is determined by a manual analysis of the relevance of the categories with respect to the subject, allowing to discard some categories that may not be relevant to the targeted subject. This strategy has been created with the aim of extracting resources related to a specific subject that are present in the graph describing a DBpedia resource. Finally the set of triples describing the resources in *R* are stored in a named graph *NG*, represented as "Larger DBpedia subgraph" in Figure 2.

Algorithm 1: Named Entity and Filtered Category Extraction (NEFCE)

NEFCE (R)
  inputs : A set of DBpedia resources R; A threshold k
  output: A named graph NG
  $C \leftarrow \emptyset$;
  foreach *resource* $r_i \in R$ **do**
    $C \leftarrow C + resource\_categories(r_i, DBpedia)$;
  $C_{topK} \leftarrow top\_categories\_with\_more\_resources(k)$;
  foreach *resource* $r_i \in R$ **do**
    $WL \leftarrow wikilink\_resources(r_i, DBpedia)$;
    foreach *wikilink resource* $wlr_i \in WL$ **do**
      $RC \leftarrow resource\_categories(wlr_i, DBpedia)$;
      foreach *resource category* $c_i \in RC$ **do**
        if *resource category* $c_i \in C_{topK}$ **then**
          $R \leftarrow R + wlr_i$;
  $NG \leftarrow generate\_named\_graph(R)$;
  return NG;

As a result, with a SPARQL query like the one shown below, it is possible to retrieve all the DBpedia resources contained in the generated named graph:

```
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT  DISTINCT ?resource ?label
FROM <http://intelliquiz/kg1>
WHERE
{
    ?resource rdfs:label ?label .
    FILTER ( strstarts(str(?a), "http://fr.dbpedia.org/resource") )
}
```

The main advantage of the *NEFCE* heuristics is that it enables to discover relevant related resources from DBpedia and it limits the number of non-relevant resources that may be added through a category-based filter. In section 6 we describe the comparative evaluation of this heuristics against the baseline (a heuristics that applies only an entity linking process to textual representations) in terms of relevance.

5 Automatic Generation of Questions from a Knowledge Graph

To automatically generate questions from a knowledge graph (generated as described in the previous section), we have implemented an existing approach (Rodriguez Rocha and Faron Zucker [2017]) which proposes a new classification, depicted in Table 1, of an existing set of strategies (Papasalouros et al. [2008]) to generate questions from a knowledge graph based on its objective. This process corresponds to the last 2 steps of Figure 2.

The first column of Table 1 lists all the categories of strategies to generate questions. As it can be seen on this table, each category corresponds to a type of targeted question

Category	Output	Example Question
Find resources of a given class	1 resource and 1 class	What is the *Seine*? a *river*
Find a data property that relates a resource and a data value	1 resource, 1 data property and 1 data value	What is the *total population* of *Paris*? *2240621*
Find an object property that relates two resources	1 property and 2 resources	In which *city* is the *Université Nice Sophia Antipolis*? in *Nice*
Find super classes of a given class	2 classes (1 class and its superclass)	What is a *Synagogue*? a *religious building*

Table 1 Categories of strategies to generate questions from a knowledge graph and their expected outputs

to be generated; and to achieve it, the expected outputs are a special kind of RDF triples that we call *Statements*, which should be retrieved from a knowledge graph by means of predefined SPARQL queries. Below is an example of a SPARQL query corresponding to a strategy of the first category "Find resources of a given class", which provides a set of resources (instances) related to their classes through the `rdf:type` property:

```
SELECT  DISTINCT ?a ?classA ?aLabel ?classALabel
WHERE
{  ?a rdf:type ?classA .
   ?a rdfs:label ?aLabel .
   ?classA rdfs:label ?classALabel .
   FILTER (strstarts(str(?classA), "http://dbpedia.org/ontology"))
   FILTER (strstarts(str(?a), "http://fr.dbpedia.org/resource"))
}
```

Finally, the questions in natural language are generated by applying category-specific string templates and NLP functions to the labels of the elements of the resulting *Statements*. Below is a string template implemented in Python, corresponding to the statements obtained from the previous example SPARQL query:

```
from string import Template
def template1(aLabel, classALabel):
    return Template('What is the $a? a $classA').safe_substitute(a=aLabel, classA=classALabel)
```

6 Empirical Validation

In this section, we describe the empirical evaluation of the proposed approach that we have carried out. This evaluation is divided in two parts. In the first one (Section 6.1), we have evaluated the proposed *NEFCE* heuristics (described in Section 4) against the baseline by comparing the relevance and the number of resources extracted for a given knowledge graph. Then, we measured the relevance and number of questions generated from the same knowledge graph.

In the second part (Section 6.2), we have comparatively evaluated the two educational referentials of knowledge and skills (*Les Incollables* and *EduProgression*) described in this article. To achieve this, by applying the proposed heuristics, we have generated a knowledge graph from the terms and textual descriptions defined by each referential for the *CM2* school year of the French educational system. After this, we have evaluated the relevance of the resources that are contained in each knowledge graph. Finally, we have measured the impact of each of these resources in terms of relevance and number of the questions generated.

6.1 Evaluation of the proposed heuristics

For this evaluation, we considered a set of 126 questions representative of a topic of geography, extracted from the Knowledge Base *Les Incollables* (described in Section 3.2). We applied to this set of questions the *NEFCE* heuristics as well as the baseline heuristics, that is a heuristics that applies only an entity linking process to the text of such questions. As a result we have generated two knowledge graphs (one for each heuristics) that contain the extracted DBpedia resources and their description.

We first evaluated the relevance of the resources selected by each heuristics and then we measured the impact that the resources generated by each heuristics can have on the generated questions.

6.1.1 Relevance of the selected resources

We asked three school teachers of Geography to evaluate the relevance of all the resources obtained from both heuristics (and merged to avoid duplicates), considering that a "relevant resource" is a resource that is related to the specified topic of Geography. A list of resources to be evaluated on a scale of 1 to 5 (where 5 is the most relevant) was provided to them in a spreadsheet.

After having evaluated the relevance of the resources, we proceeded to calculate the precision and the recall of the baseline and of our heuristics. We defined them as the proportion of relevant resources among all the resources generated by a given heuristics and the proportion of relevant resources generated by a given heuristics among all the relevant resources generated by any of the two heuristics, respectively. According to the previous defined scale of relevancy, we considered that a resource is sufficiently relevant if its score is greater than or equal to 3 out of 5. The average precision and recall (considering the three evaluators) are reported in Table 2 and shown in Figure 3.

	Baseline	**NEFCE**
Extracted resources	99	318
Precision	0,95	0,87
Recall	0,34	1

Table 2 Number of extracted resources, precision and recall per heuristics

The results show that the proposed heuristics obtains a much higher recall than the baseline while loosing very little in precision.

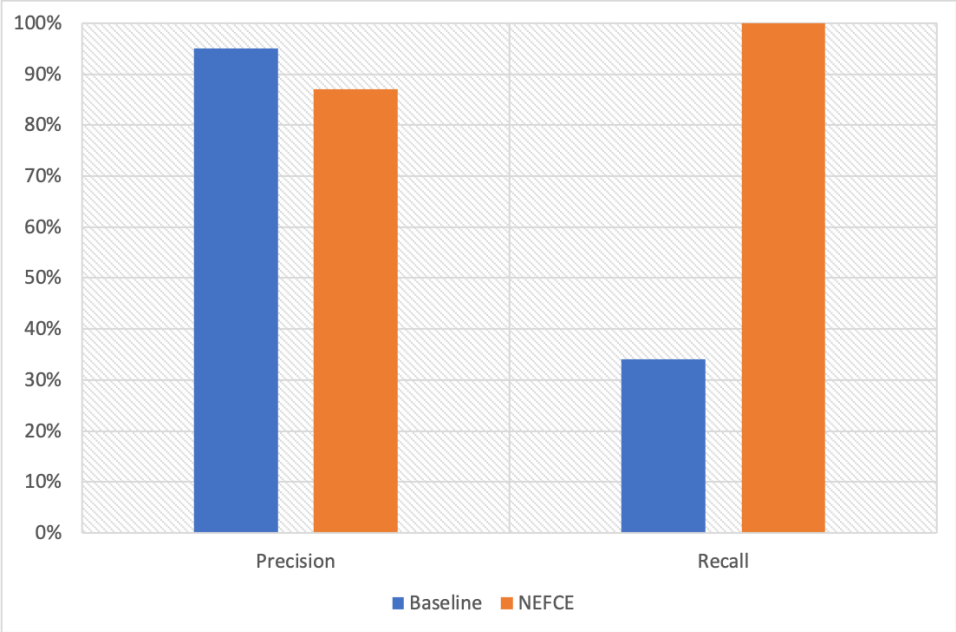


Figure 3: Precision and recall of each heuristics

6.1.2 Relevance of the generated questions

After conducting the evaluation of relevance of the resources selected by each heuristics, we applied the techniques proposed in (Rodriguez Rocha and Faron Zucker [2017]) for the automatic generation of quizzes to each DBpedia subgraphs resulting from the *NEFCE* heuristics and the baseline. Then, we asked the evaluators to evaluate the relevance of the generated questions according to the Geography domain (on a scale of 1 to 5, where 5 is the most relevant). For this, the evaluators have been been provided with a list of 100 questions, randomly extracted from each subgraph.

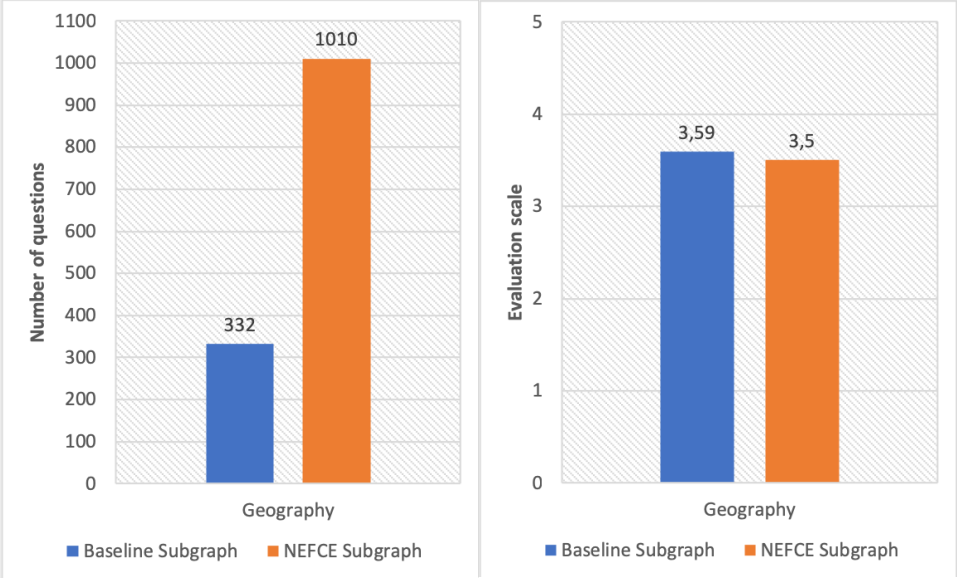
The results of this evaluation (Figure 4) show that the proposed heuristics enables to generate a much higher number of questions while loosing very little in relevancy.

6.2 Evaluation of the educational references

To perform the comparative evaluation of (*Les Incollables* and *EduProgression*), the two referentials of knowledge and skills, we generated a knowledge graph for each referential from the terms and the textual descriptions that they provide for the *CM2* school year of the French educational system. The evaluations of the relevance of the resources and the generated questions from each knowledge graph are reported below.

6.2.1 Evaluation of the selected resources

In order to assess the relevance of the resources contained in the two knowledge graphs for the CM2 school year in Geography, we asked a geography teacher who teaches the CM2 level, to evaluate the relevance of a list of resources to Geography and CM2. The teacher was provided with a spreadsheet that contained a list of 188 different resources, resulting



(a) Number of questions per heuristics (b) Question's relevance per heuristics

Figure 4: (a) Number and (b) average relevance of the questions per heuristics

from the random extraction of 100 resources from each knowledge base and merged to remove the duplicates. Each resource was evaluated on a scale of 1 to 5, where 1 meant that the resource was "not relevant at all" and 5 that the resource was "very relevant". Once the relevance of the resources was evaluated, we calculated the precision obtained for each referential, defined as the proportion of relevant resources among the set of resources generated with the considered referential (equation (1)). Resources whose score was equal to or greater than 3 were considered as "relevant".

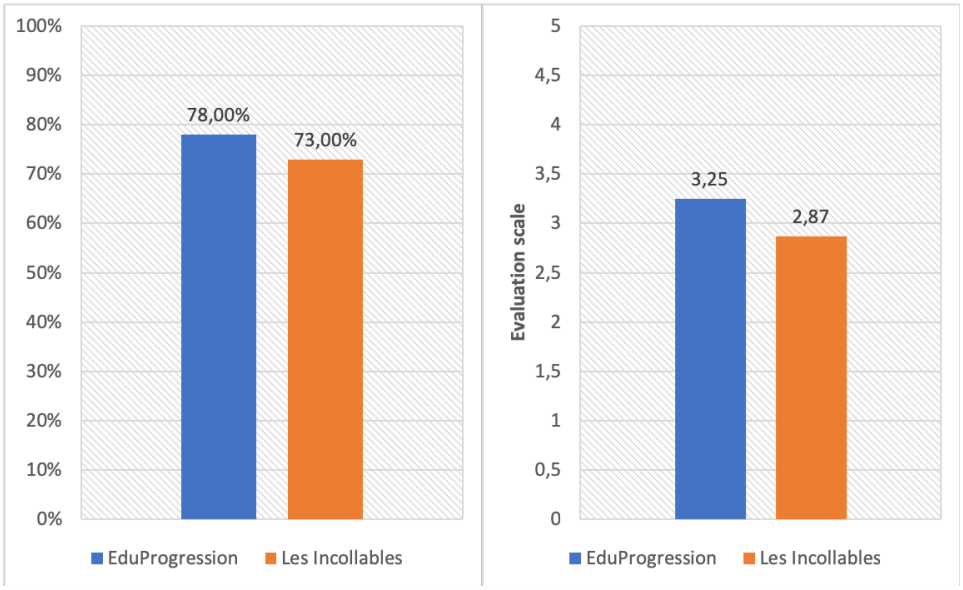
$$P = \frac{Number\ of\ Relevant\ Resources}{Total\ Set\ of\ Resources\ Generated\ from\ a\ Referential} \tag{1}$$

The results are provided in Figure 5a. They show that *EduProgression* slightly outperforms *Les Incollables*: the DBpedia resources extracted with *EduProgression* are slightly more relevant according to the domain and to the school level than those extracted with *Les Incollables*.

6.3 Evaluation of the generated questions

Once the pertinence of the DBpedia resources extracted with the two educational referentials has been evaluated, we have applied the quiz generation techniques mentioned in Section 5, to the DBpedia subgraph generated from each referential. We generated 191 and 286 questions from the *EduProgression* and the *Les Incollables* subgraphs respectively.

Finally, we have asked the geography teacher to evaluate the relevance of the questions generated from each educational referential (on a scale of 1 to 5, where 5 is the most relevant). For this, he has been given a list of 100 questions for each educational reference, extracted randomly.



(a) Resource's precision (b) Question's average relevance

Figure 5: (a) Precision of the resources and (b) average relevance of the questions per educational reference of knowledge and skills

Below, the Figure 5b shows the results of this evaluation. They are consistent with the above assessed relevance of the evaluated resources extracted from each educational reference, showing that *EduProgression* outperforms *Les Incollables*: the questions generated from *EduProgression* are more relevant according to the domain and to the school level than those generated from *Les Incollables*.

7 Conclusions and Future Work

We have presented an approach to automatically generate questions related to a specific subject and oriented to the school level of the target student. Considering an educational reference, our approach enables to extract resources from DBpedia relevant for a specific subject and a specific school year, from which questions are generated. Two structured educational references: *EduProgression* and *Les Incollables* have been presented and evaluated in order to determine from which one it is possible to extract more relevant resources to allow to generate more relevant questions. From the evaluation of relevance made to the resources extracted from DBpedia corresponding to the CM2 school year of geography, and to the questions generated from them, we were able to conclude that *EduProgression* was the educational reference that allowed to extract resources and generate questions with greater relevance.

We have also proposed and detailed *NEFCE*, a heuristics to extract relevant DBpedia resources. The experiments carried out and described in this research work have shown that (1) The enrichment of the knowledge graph that contains semantically related DBpedia resources, makes it possible to increase the number of generated questions, and (2) ranking

the candidate related resources by their degree of relevancy to a domain, enables to maintain the precision of the generation of questions.

As future work, we plan to extend the evaluation of our automatic quiz generation approach by evaluating generated quizzes from different domains and school courses and also by involving more professors in each evaluation. Additionally, the generation of automatic quizzes has been done using references focused on French education. It is very important in the future to carry out experiments with educational references from other countries.

For the moment our approach relies on the extraction of resources from DBpedia, a dataset that contains structured data extracted from Wikipedia, an encyclopedia. On the one hand this is an advantage as we can generate quizzes with generic questions about many different subjects. On the other hand this can be an inconvenience when it is required to generate quizzes with very specialized questions about a specific domain. For this we plan to add support for generating questions also from other linked data sources and / or domain ontologies.

Finally we believe that it is of prime importance to add to our approach a step in which the teacher can have the possibility to curate manually the extracted ressources and generated questions, that is, remove those that are not relevant and edit their text when necessary.

References

Maha Al Yahya, Remya George, and Auhood Alfaries. Ontologies in e-learning: Review of the literature. *International Journal of Software Engineering and Its Applications*, 9 (2):67–84, 2015.

Maha Al-Yahya, Auhood Al-Faries, and Remya George. Curonto: An ontological model for curriculum representation. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '13, pages 358–358, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2078-8. doi: 10.1145/2462476.2465602. URL `http://doi.acm.org/10.1145/2462476.2465602`.

Hilary Dexter and Ioan Davies. An ontology-based curriculum knowledgebase for managing complexity and change. *2014 IEEE 14th International Conference on Advanced Learning Technologies*, 0:136–140, 2009. doi: http://doi.ieeecomputersociety.org/10.1109/ICALT.2009.85.

M. Bucos, B. Dragulescu, and M. Veltan. Designing a semantic web ontology for e-learning in higher education. In *Electronics and Telecommunications (ISETC), 2010 9th International Symposium on*, pages 415–418, Nov 2010. doi: 10.1109/ISETC.2010.5679298.

Cui Guangzuo, Chen Fei, Chen Hu, and Li Shufang. Ontoedu: a case study of ontology-based education grid system for e-learning. In *GCCCE International conference, Hong Kong*, 2004.

Oscar Rodríguez Rocha, Catherine Faron Zucker, and Geraud Fokou Pelap. A formalization of the french elementary school curricula. In Paolo Ciancarini, Francesco Poggi, Matthew Horridge, Jun Zhao, Tudor Groza, Mari Carmen Suarez-Figueroa, Mathieu d'Aquin, and Valentina Presutti, editors, *Knowledge Engineering and Knowledge Management*, pages 82–94, Cham, 2017. Springer International Publishing. ISBN 978-3-319-58694-6.

S. Lalithsena, P. Kapanipathi, and A. Sheth. Harnessing relationships for domain-specific subgraph extraction: A recommendation use case. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 706–715, 2016. doi: 10.1109/BigData.2016.7840663.

Roberto Mirizzi, Azzurra Ragone, Tommaso Di Noia, and Eugenio Di Sciascio. Semantic tags generation and retrieval for online advertising. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1089–1098, New York, NY, USA, 2010a. ACM. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871576. URL http://doi.acm.org/10.1145/1871437.1871576.

Roberto Mirizzi, Azzurra Ragone, Tommaso Di Noia, and Eugenio Di Sciascio. Semantic tag cloud generation via dbpedia. In Francesco Buccafurri and Giovanni Semeraro, editors, *E-Commerce and Web Technologies*, pages 36–48, 2010b. ISBN 978-3-642-15208-5.

Roberto Mirizzi, Azzurra Ragone, Tommaso Di Noia, and Eugenio Di Sciascio. Semantic wonder cloud: Exploratory search in dbpedia. In Florian Daniel and Federico Michele Facca, editors, *Current Trends in Web Engineering*, pages 138–149, 2010c. ISBN 978-3-642-16985-4.

Cristhian Figueroa, Iacopo Vagliano, Oscar Rodríguez Rocha, and Maurizio Morisio. A systematic literature review of linked data-based recommender systems. *Concurrency and Computation: Practice and Experience*, 2015. ISSN 1532-0634. doi: 10.1002/cpe.3449. URL http://dx.doi.org/10.1002/cpe.3449.

Marija Cubric and Milorad Tosic. Towards automatic generation of e-assessment using semantic web technologies. In *Proceedings of the 2010 International Computer Assisted Assessment Conference*, 2010. URL http://hdl.handle.net/2299/4885.

M. Al-Yahya. Ontoque: A question generation engine for educational assesment based on domain ontologies. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*, pages 393–395, July 2011. doi: 10.1109/ICALT.2011.124.

Muriel Foulonneau. *Generating Educational Assessment Items from Linked Open Data: The Case of DBpedia*, pages 16–27. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-25953-1. doi: 10.1007/978-3-642-25953-1_2. URL http://dx.doi.org/10.1007/978-3-642-25953-1_2.

Asma Ben Abacha, Júlio Cesar dos Reis, Yassine Mrabet, Cédric Pruski, and Marcos Da Silveira. Towards natural language question generation for the validation of ontologies and mappings. *J. Biomedical Semantics*, 7:48, 2016. doi: 10.1186/s13326-016-0089-6. URL http://dx.doi.org/10.1186/s13326-016-0089-6.

Naveed Afzal and Ruslan Mitkov. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, 18(7):1269–1281, 2014. ISSN 1433-7479. doi: 10.1007/s00500-013-1141-4. URL http://dx.doi.org/10.1007/s00500-013-1141-4.

Tahani Alsubait, Bijan Parsia, and Uli Sattler. Generating multiple choice questions from ontologies: Lessons learnt. In *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014) co-located with 13th International Semantic Web Conference on (ISWC 2014), Riva del Garda, Italy, October 17-18, 2014.*,

pages 73–84, 2014. URL http://ceur-ws.org/Vol-1265/owled2014_submission_11.pdf.

Tahani Alsubait, Bijan Parsia, and Uli Sattler. *Generating Multiple Choice Questions From Ontologies: How Far Can We Go?*, pages 66–79. Springer International Publishing, Cham, 2015. ISBN 978-3-319-17966-7. doi: 10.1007/978-3-319-17966-7_7. URL http://dx.doi.org/10.1007/978-3-319-17966-7_7.

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. Ontology-based multiple choice question generation. *KI - Künstliche Intelligenz*, 30(2):183–188, 2016. ISSN 1610-1987. doi: 10.1007/s13218-015-0405-9. URL http://dx.doi.org/10.1007/s13218-015-0405-9.

Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos Kotis. Automatic generation of multiple choice questions from domain ontologies. In Miguel Baptista Nunes and Maggie McPherson, editors, *e-Learning*, pages 427–434. IADIS, 2008. ISBN 978-972-8924-58-4. URL http://dblp.uni-trier.de/db/conf/iadis/el2008.html#PapasalourosKK08.

O. Rodriguez Rocha and C. Faron Zucker. Automatic generation of educational quizzes from domain ontologies. In *EDULEARN17 Proceedings*, 9th International Conference on Education and New Learning Technologies, pages 4024–4030. IATED, 2017. ISBN 978-84-697-3777-4. doi: 10.21125/edulearn.2017.1866. URL http://dx.doi.org/10.21125/edulearn.2017.1866.

Note

[1] http://media.education.gouv.fr/file/46/7/5467.pdf
[2] http://ns.inria.fr/semed/eduprogression
[3] http://www.inria.fr
[4] http://www.qwant.com
[5] http://www.lesincollables.com
[6] https://en.wikipedia.org/wiki/Optical_character_recognition
[7] http://ns.inria.fr/semed/linquest
[8] http://www.gaya-technology.com/incollables
[9] http://ns.inria.fr/semed/frenchedu/
[10] https://en.wikipedia.org/wiki/Entity_linking
[11] http://www.dbpedia-spotlight.org/